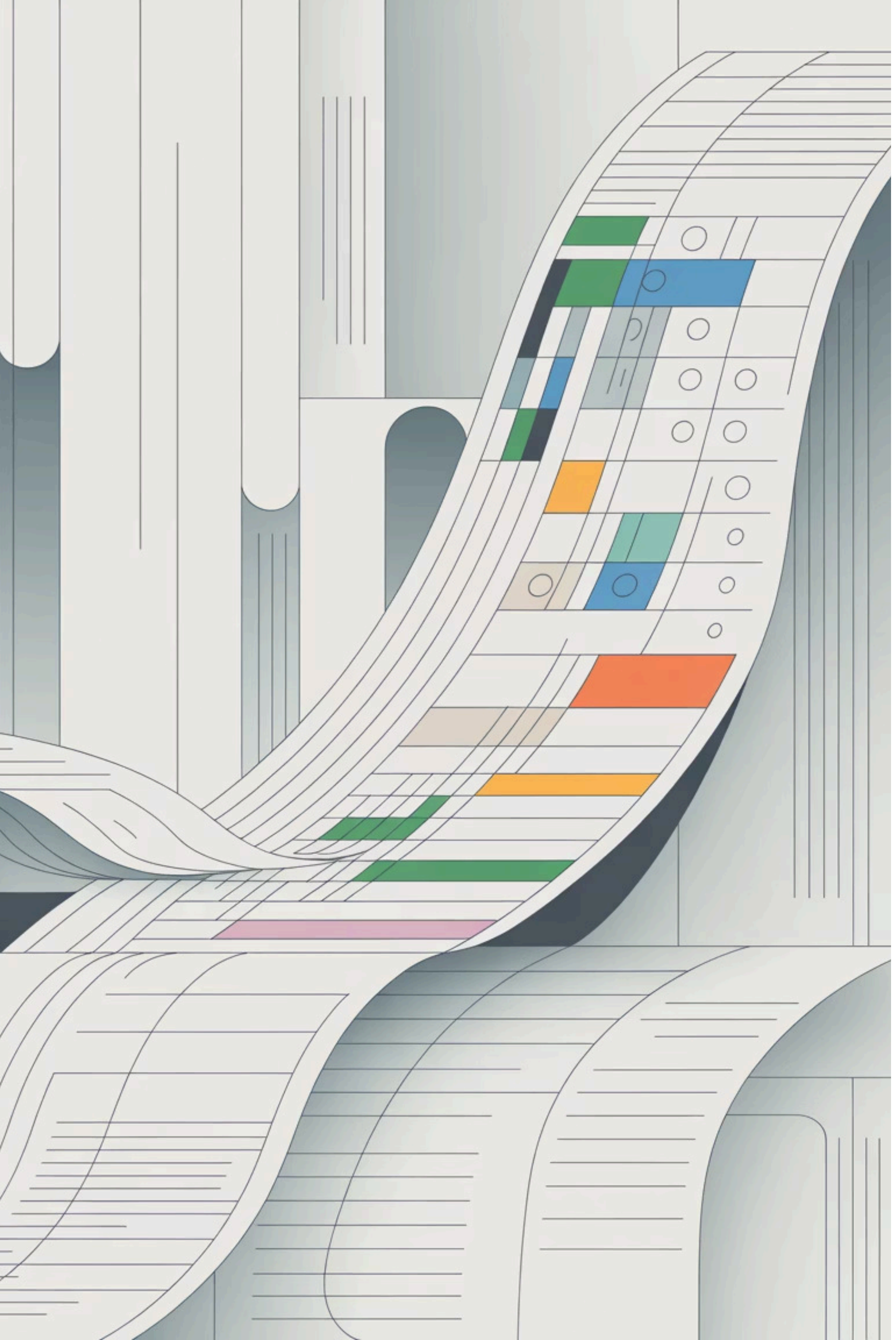
Applied NLP

Session 2

Lecturer: Narges Chinichian

Winter Semester 2025-2026



Session 2: Analyzing Phrases in Text

Moving beyond individual words to explore multi-word structures and their significance in computational text analysis.

Quick Recap: What We've Learned

Frequency Analysis

Counted word occurrences to identify patterns

Adverbs & Modifiers

Examined how authors emphasize and qualify

Punctuation Patterns

Analyzed stylistic markers in writing

Semantic Categories

Explored gendered language and color vocabulary

These foundational word-level measures set the stage for today's deeper exploration of how words combine to create meaning.

Why Phrases Matter

Beyond Single Words

Individual words tell part of the story, but phrases reveal an author's unique fingerprint. Collocations—words that habitually appear together—create distinctive patterns that define style and meaning.

Idioms and fixed expressions carry cultural weight that transcends their component words.

The Power of Collocation

> "Strong tea" feels natural and expected.
>
> "Powerful tea" sounds unusual and marked.

These preferences aren't random—they're learned patterns that computational methods can measure and compare across texts.

The Linguistic Unit: Phrases

Bigrams

Two consecutive words forming a unit (e.g., "dark forest," "she said")

Tri/Ngrams

Three/N-word sequences capturing longer patterns (e.g., "at the end," "oh oh oh", "poor little thing")

Idiomatic Patterns

Fixed expressions with specialized meanings beyond literal interpretation

Authorial Signatures

Recurring multi-word combinations that distinguish one writer from another

Level 2: The Phrase Level

1 **Word**

Individual lexical items—our foundation from Session 1

2 **Phrase**

Today's focus: Multi-word units and collocation patterns

3 **Sentence**

Coming next: Syntactic structures and complexity measures

4 **Paragraph**

Discourse organization and coherence analysis

5 **Text**

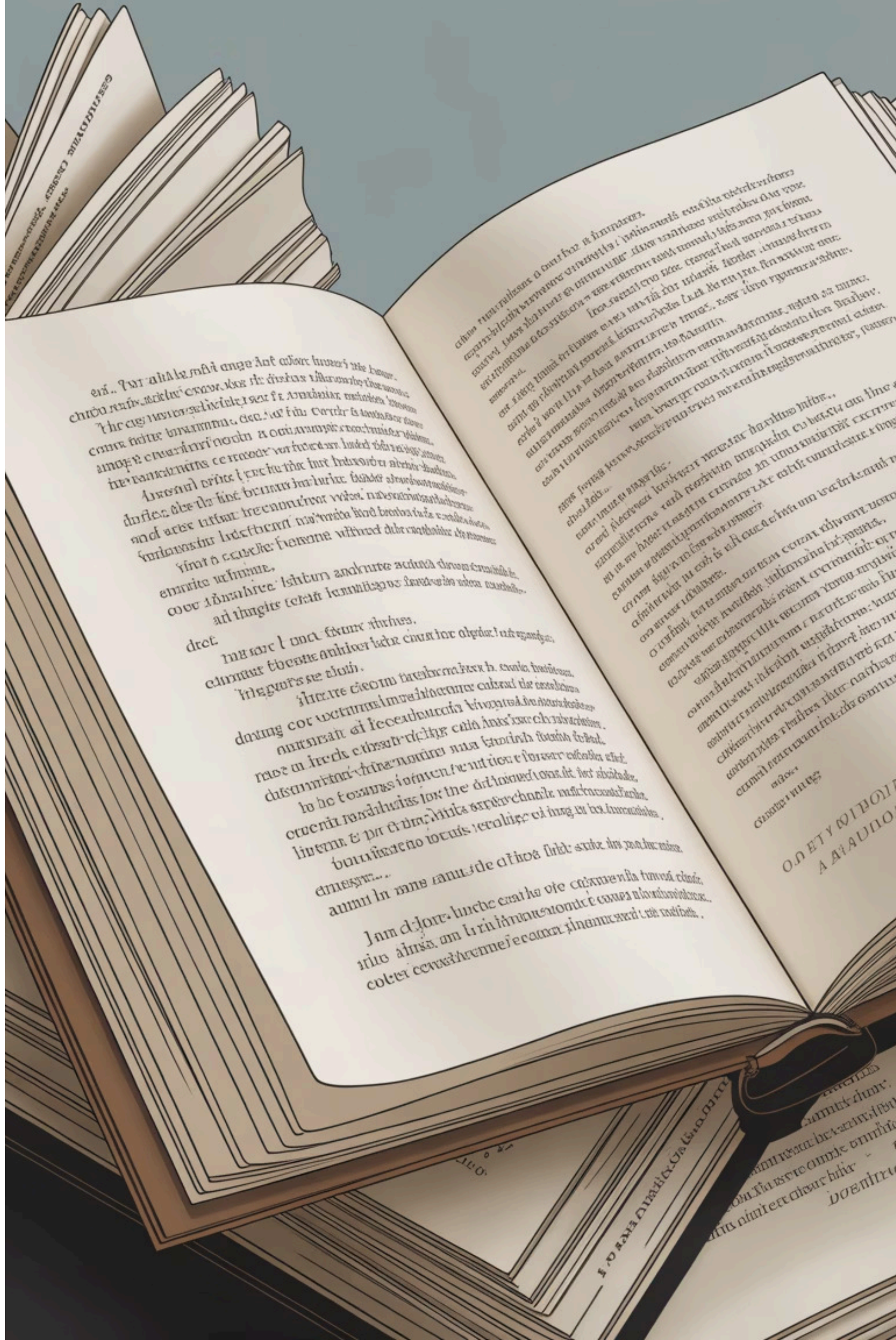
Document-level patterns and macrostructure

Each level builds on the previous, creating a comprehensive framework for computational text analysis.

In the Alice books by Lewis Carroll

- **“Oh oh oh”** and **“poor little thing”**: recurring exclamations that convey emotion and empathy.
- **“thought to herself”** (Through the Looking-Glass) and **“said to herself”** (Wonderland) : mark Alice’s inner reflections and spoken self-dialogue.
- **“she went on”**: signals Alice’s over-talkative curiosity and logical rambling.

These frequent multi-word units show how Carroll’s patterned language builds a distinctive narrative voice—playful, self-aware, and rhythmically reflective.



Measure 1: N-gram Frequency

1

Extract N-grams

Use sliding window to capture all consecutive word pairs (bigrams) or triples (trigrams) from the text

2

Count Occurrences

Tally how many times each unique n-gram appears across the entire corpus

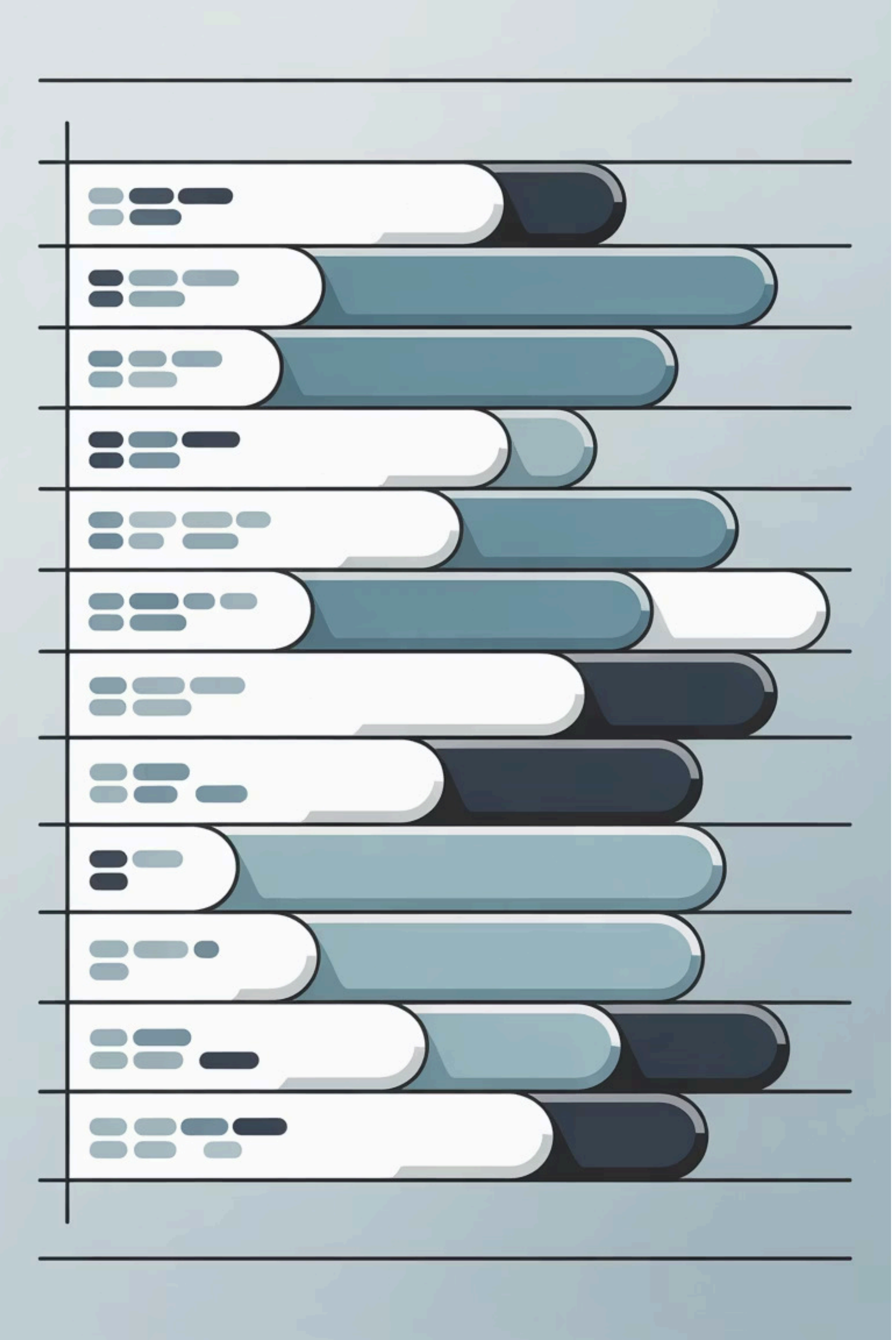
3

Rank & Filter

Sort by frequency and optionally remove stopword-only phrases to surface meaningful patterns

Formula: For a text of N words, there are (N-1) bigrams and (N-2) trigrams to analyze.

This straightforward counting method reveals which multi-word expressions authors favor, providing insight into their habitual language patterns.



Demo: Top N-grams Visualized

Notebook 1: Bigram & Trigram Frequency

The bar chart displays the most frequent bigrams and trigrams from our corpus. Notice how certain phrases dominate the rankings.

Key Observations

- Function words create high-frequency bigrams
- Content-rich trigrams appear less frequently
- Genre influences phrase distribution
- Preprocessing decisions are essential

Filtering Strategies

Remove pure stopword sequences to uncover meaningful collocations. Set minimum frequency thresholds to reduce noise while preserving significant patterns.

Measure 2: Pointwise Mutual Information

The PMI Formula

PMI measures how much more often two words appear together than we'd expect by chance, based on their individual frequencies.

High PMI = Strong Association

When PMI is positive and large, the collocation is meaningful—words prefer each other's company.

$$\mathrm{PMI}(x, y) = \log_2 \left(\frac{P(x, y)}{P(x) \times P(y)} \right)$$

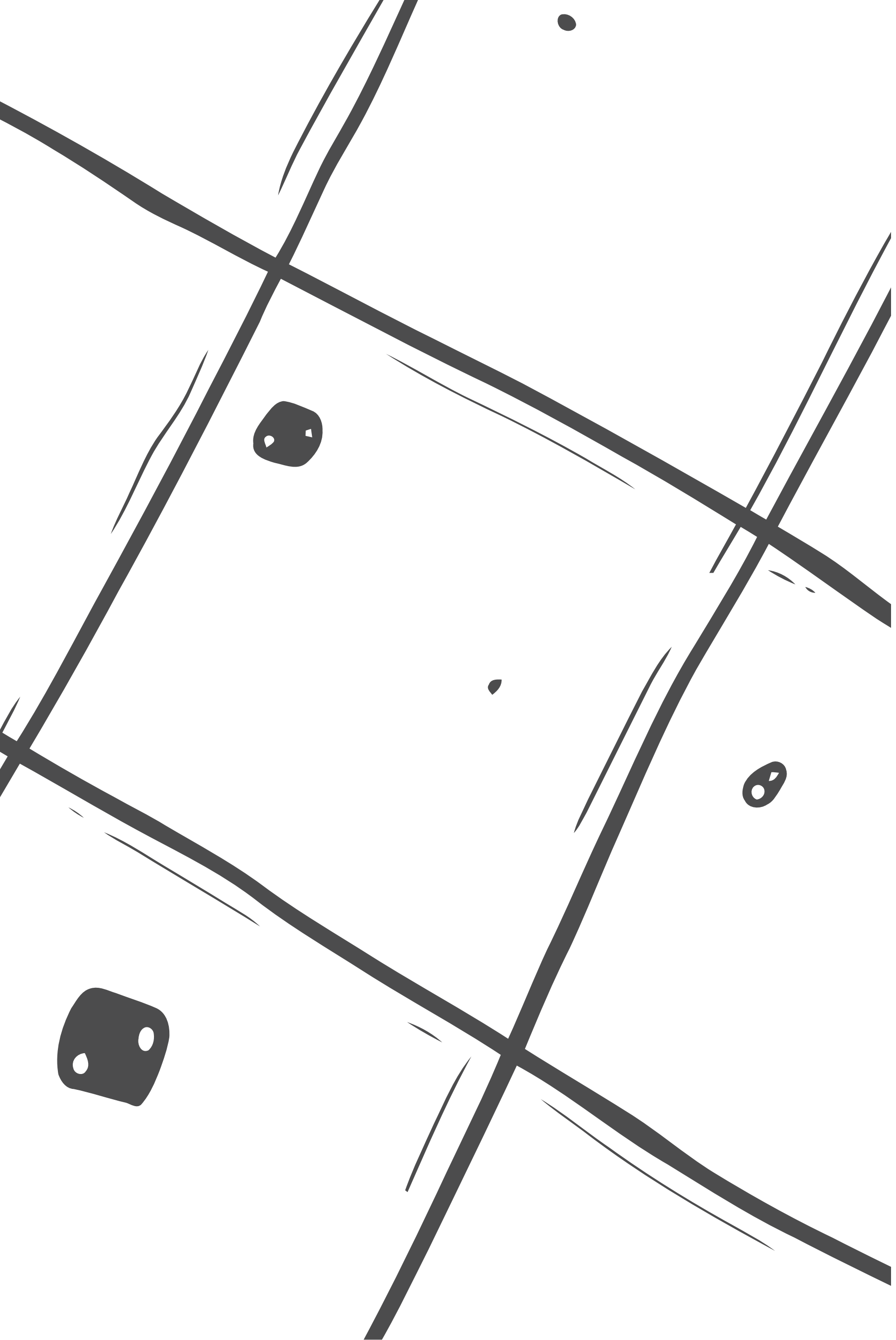
Where:

- $P(x, y)$ is the probability of x and y occurring together
- $P(x)$ is the probability of x occurring
- $P(y)$ is the probability of y occurring

Example: Color Collocations

"Blood red" — High PMI
These words strongly associate; the combination carries specific semantic weight.

"Red apple" — Lower PMI
While common, the association is weaker because "red" and "apple" appear with many other words.



Demo: PMI Collocations

Notebook 2: Discovering Strong Associations

The table shows top-scoring collocations after applying PMI calculations with a minimum frequency filter. This filter prevents rare coincidences from inflating scores.

Frequency Threshold

Set minimum occurrence count (e.g., 5+ times) to ensure collocations are genuine patterns, not statistical flukes

Interpretation

Higher PMI scores indicate tighter lexical bonds—phrases that function as semantic units in the author's vocabulary

Measure 3: Part-of-Speech Pattern Frequency

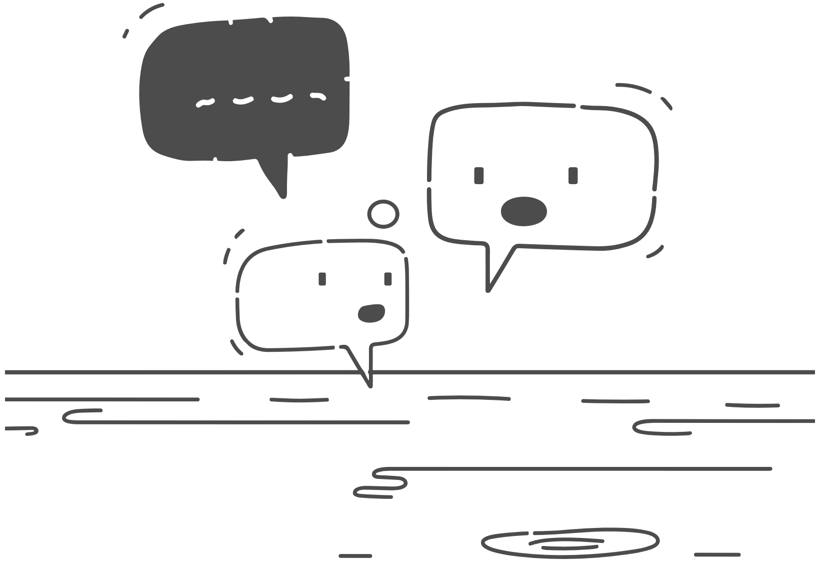
Beyond Surface Forms

Instead of counting specific word sequences, we extract grammatical templates that reveal syntactic preferences.

An author might favor adjective-noun pairings or verb-adverb combinations regardless of the particular words used.

Common POS Patterns

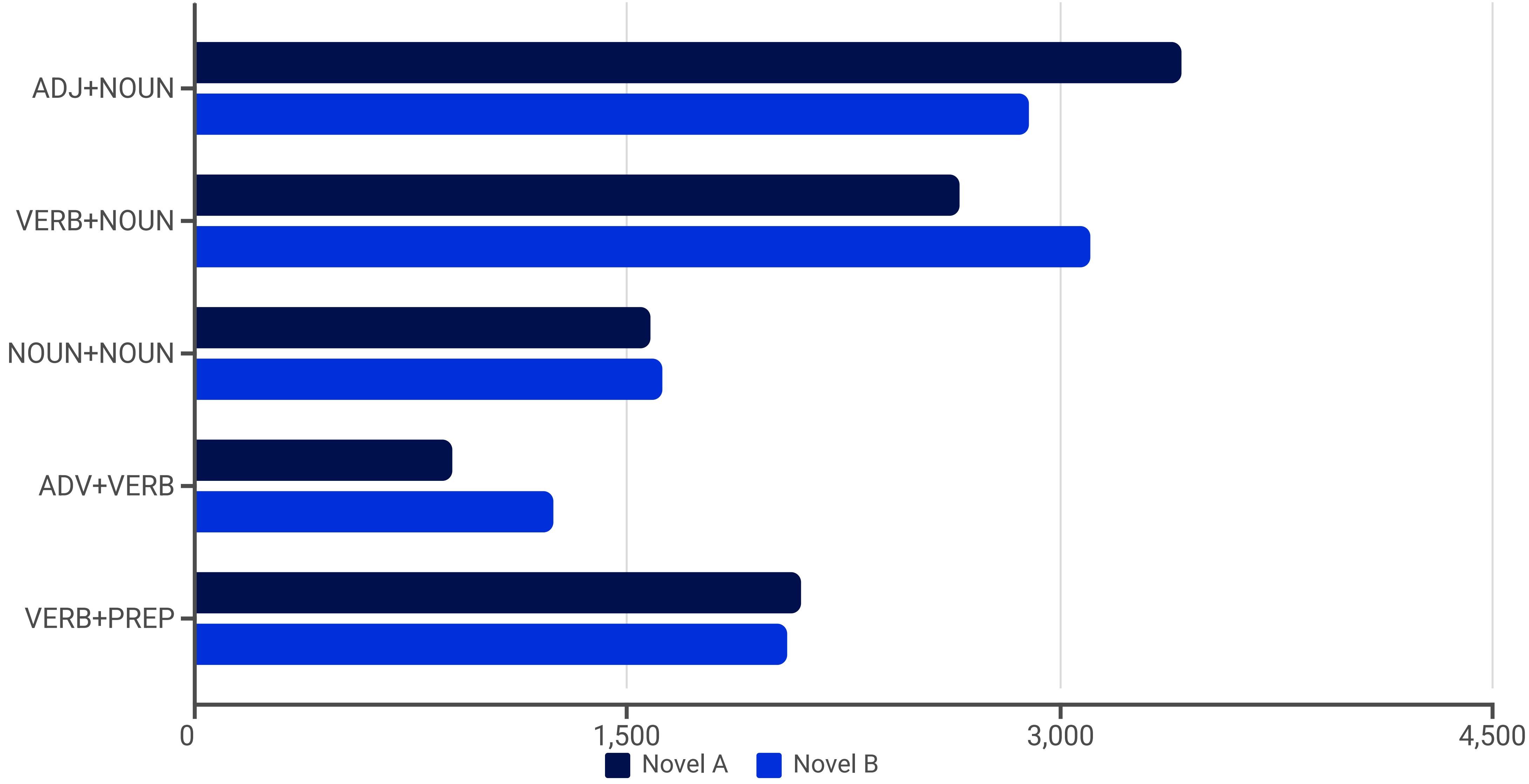
- **Adjective + Noun** → "dark forest," "happy child"
- **Verb + Noun** → "broke silence," "felt pain"
- **Noun + Noun** → "kitchen table," "summer afternoon"
- **Adverb + Verb** → "quietly entered," "quickly realized"
- **Verb + Preposition** → "looked at," "walked through"



These abstractions help us compare authors at a grammatical level, identifying whether one writer prefers modification over action, or nominal versus verbal style.

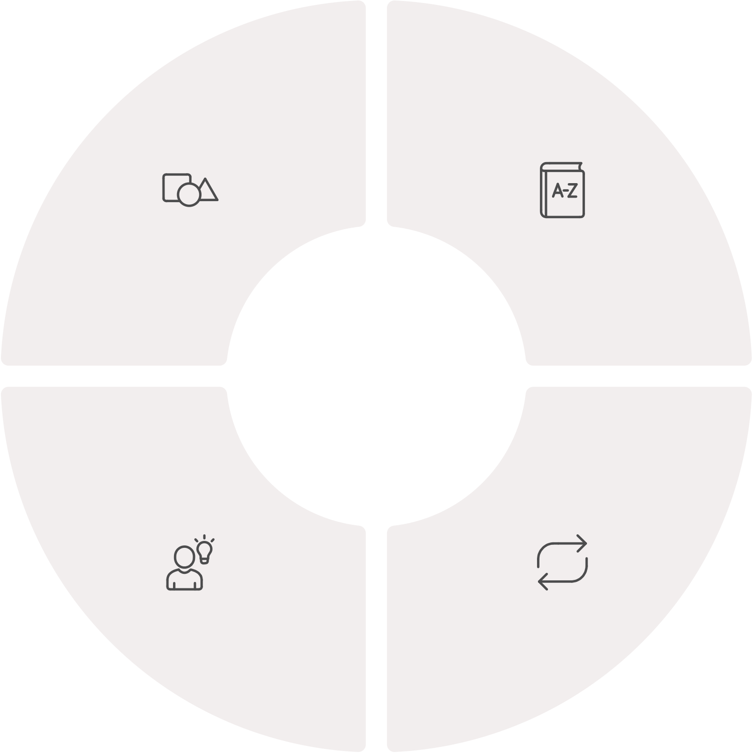
Demo: Comparing POS Patterns

Notebook 3: Same Author, Different Works



Notice how the same author shows consistency in some patterns while varying others—perhaps reflecting different genres, periods, or narrative perspectives across their works.

Measure 4: Phrase Diversity



Type-Token Ratio

Calculate unique phrase types divided by total phrase tokens

Lexical Richness

Higher ratios indicate greater variety in multi-word expressions

Creative Range

Links to authorial originality and linguistic inventiveness

Repetition Index

Lower diversity suggests formulaic or repetitive phrasing

Just as we measured lexical diversity for words, phrase diversity reveals whether an author employs a wide range of multi-word expressions or relies on familiar combinations.

Demo: Phrase Diversity Across Books

Notebook 4: Tracking Variation

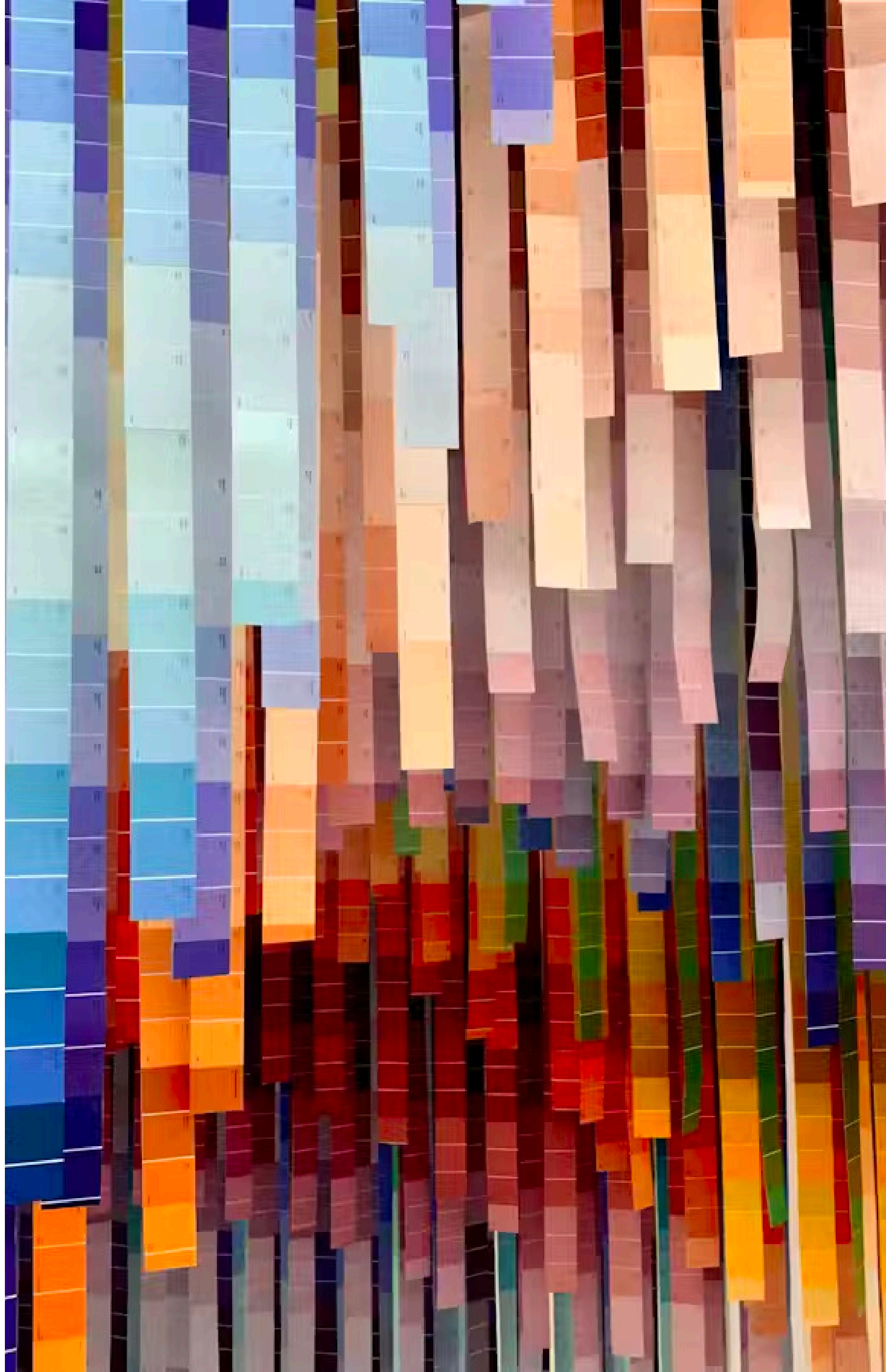
Interpreting the Plot

Each number represents the type-token ratio for bigrams in one book. Higher values indicate books with greater phrasal variety; lower values suggest more repetitive language.

Possible Explanations

- Dialogue-heavy **books** may show higher diversity
- Books with more action sequences might repeat formulaic phrases
- Differences in authorial style or genre can significantly impact overall phrase diversity

This measure helps identify how different authors or genres compare in linguistic freshness, revealing tendencies to either maintain varied phrasing or rely on established patterns—an insight valuable for stylistic analysis.



Measure 5: Collocation Networks

Visualizing Word Relationships

1

Build the Graph

Words become nodes; edges represent co-occurrence within a window

2

Weight Connections

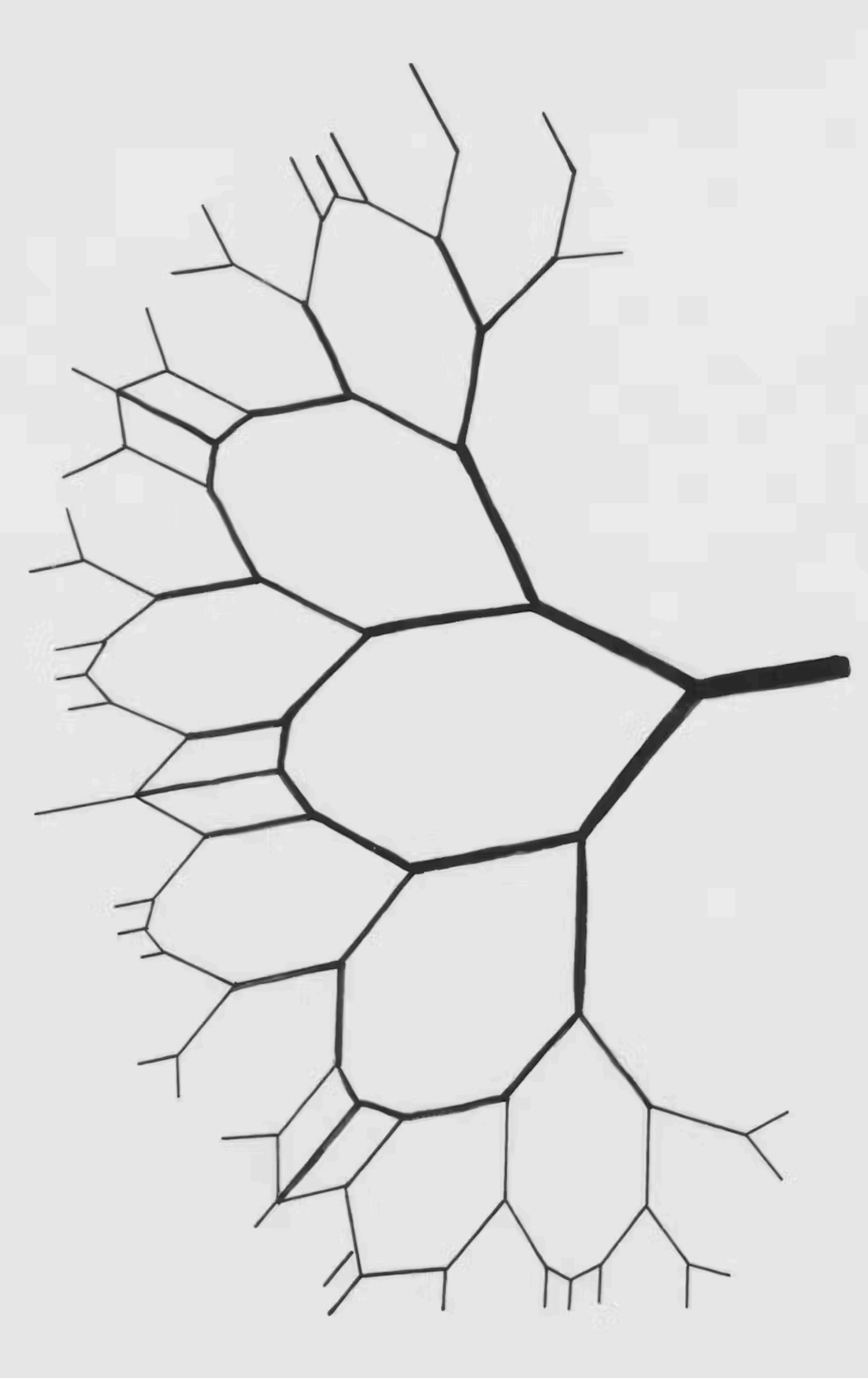
Edge thickness reflects frequency or PMI strength

3

Identify Clusters

Dense regions reveal semantic fields and thematic networks

Network visualization transforms collocation statistics into intuitive spatial representations, making it easier to spot central concepts and peripheral associations in a text.



Demo: Network Example

Notebook 5: Collocation Graph for Alice's Adventures

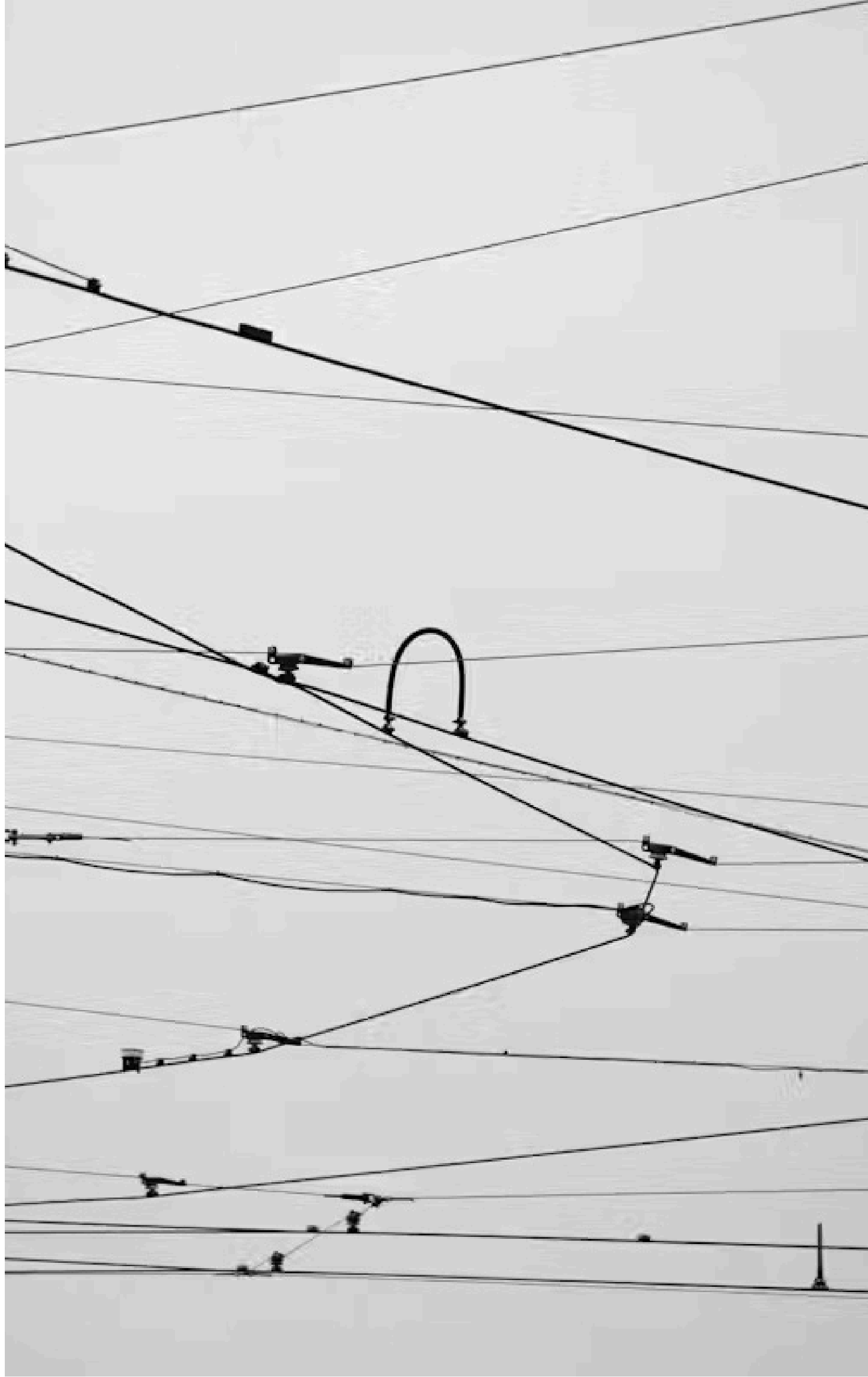
Central Nodes

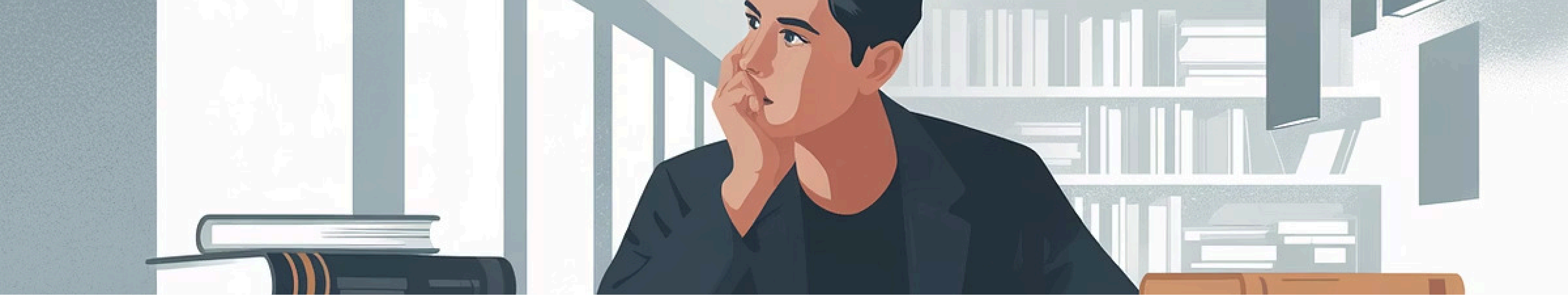
High-degree nodes such as "Alice," "said," and "thought" form the core of the network, highlighting Alice's central role and the prevalence of dialogue and introspection in her adventures. These words co-occur with many others.

Thematic Clusters

Distinct thematic clusters emerge: one for characters (e.g., Mad Hatter, Queen of Hearts, Cheshire Cat), another for Wonderland locations (e.g., tea party, garden, court), and a third for fantasy and conceptual elements (e.g., curious, wonder, dream). These groupings effectively map the key elements and narrative threads of the Alice books.

Try this: Compare networks from different authors or genres to see how vocabulary and collocation structures differ dramatically across texts.





Reflect & Compare

Author Fingerprints

What recurring phrases define your chosen author's style? Do they favor certain grammatical patterns or lexical combinations?

Genre Conventions

How do phrase patterns differ between romance novels and hard-boiled detective fiction? Between poetry and prose?

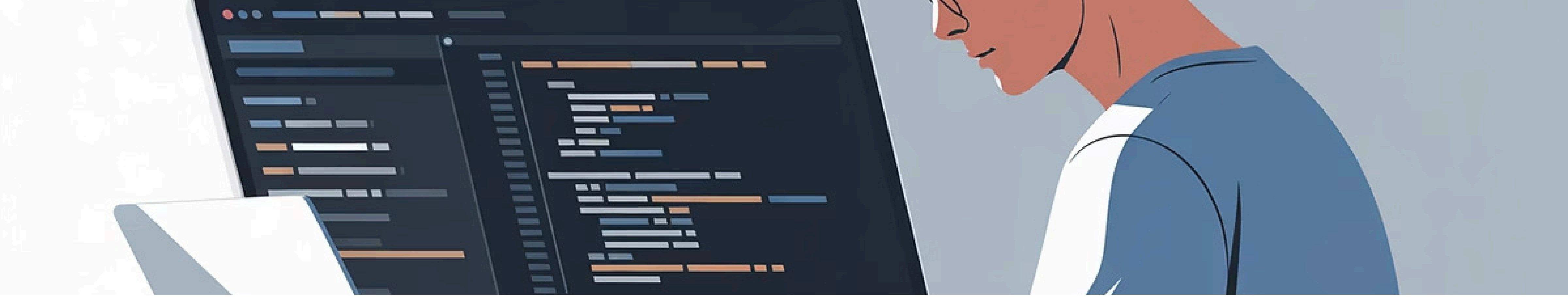
Historical Shifts

Can you detect changes in an author's phraseology across their career? Do early works show different collocation preferences than later ones?

Cultural Context

What do idiomatic expressions reveal about the cultural moment in which a text was written?

Use these questions to guide your exploration. The computational measures we've covered provide evidence for literary and linguistic arguments.



Homework Assignment

Implementation Task

Apply at least three of five phrase-level measures to your selected corpus:

1. N-gram frequency analysis
2. PMI collocation scoring
3. POS pattern extraction
4. Phrase diversity calculation
5. Collocation network visualization

Reflection Component

Write a **one-page reflection** in a notebook addressing:

- Which measures revealed unexpected patterns?
- How do results compare across texts or authors?
- What challenges did you encounter?
- What questions emerged from the analysis?

Due date: Prepare a presentation for the next session. Include visualizations for at least three of the five measures.

Looking Ahead

Session 1: Words

Frequency, categories, distribution

Session 2: Phrases

Collocations, patterns, networks

Session 3: Sentences

Syntax & complexity ahead!

Next Time: Sentences & Syntax

We'll examine how sentence length, clause structure, and syntactic complexity vary across texts. You'll learn to measure readability, parse trees, and dependency relations—moving from local collocations to global grammatical architecture.

Prepare by reviewing basic syntax concepts: clauses, phrases, and sentence types. We'll build on everything you've learned about words and multi-word units.

Questions?

Let's discuss any challenges from today's measures or clarify concepts before you begin your homework implementation.